Statistical Bias Correction Modelling for Seasonal Rainfall Forecast for the case of Bali Island

D Lealdi¹, S Nurdianti¹, A Sopaheluwakan²

¹Department of Mathematics, Bogor Agricultural University, Dramaga Raya Street, West Java 16680, Indonesia

²Indonesian Agency for Meteorology, Climatology and Geophysics, 2 Angkasa 1 Street, Kemayoran, Jakarta 10720, Indonesia

Email : dedi_math07@yahoo.co.id, nurdianti@ipb.ac.id, ardhasena@bmkg.go.id

Abstract. Rainfall is an element of climate which is highly influential to the agricultural sector. Rain pattern and distribution highly determines the sustainability of agricultural activities. Therefore, information on rainfall is very useful for agriculture sector and farmers in anticipating the possibility of extreme events which often cause failures of agricultural production. This research aims to identify the biases from seasonal forecast products from ECMWF (European Centre for Medium-Range Weather Forecasts) rainfall forecast and to build a transfer function in order to correct the distribution biases as a new prediction model using quantile mapping approach. We apply this approach to the case of Bali Island, and as a result, the use of bias correction methods in correcting systematic biases from the model gives better results. The new prediction model obtained with this approach is better than ever. We found generally that during rainy season, the bias correction approach performs better than in dry season.

Keywords: Bias Correction, Climate, ECMWF, Quantile Mapping, Rainfall.

1. Introduction

The rainfall in the tropical area is one of the factors in climate with the highest variability. Indonesia is one of the countries with a tropical climate which is geographically located between 6° North Latitude and 11° South Latitude and between 95° West Longitude and 141° East Longitude crossing the equator. The characteristics of rainfall in several regions are definitely not the same. This condition is caused by several factors, namely position of region, condition of earth in a region, mountain and valley in a region, and even structure and orientation of island. As the result, the pattern of rainfall distribution tends to be uneven among the regions in the wide scope [1].

The rainfall is one of the factors in climate which is highly influential on the agricultural sector. The consistency in rain pattern and distribution highly determines the sustainability of agricultural activities. However, this condition will be disturbed during the phenomenon of extreme climates, such as the dry extreme (El Nino) and wet extreme (La Nina). The phenomenon of extreme climate can cause rainfall pattern which deviates from the normal condition.

According to the analysis of Indonesian Agency for Meteorological, Climatological and Geophysics, in Banten, Special Capital City District of Jakarta, and West Java, the possibility of extreme rain reached 500 mm/month in 1970-1999. This possibility increased to 13% in 1900-1929, the possibility of extreme rain in these three regions is only 3% [2]. Indonesian people also experienced the real climate change in

2010. In wet drought throughout 2010, this year was ended with the drought at last, it is the real description of how the climate has really changed [3].

Based on data released by BPS Bali Province in 2014, agricultural production decreased 2.74% to 857.944 tons of dry milled grain. The decline is generally due to a decrease of paddy harvest area to 5.11% or 7,683 ha due to drought as a result of the dry season. From June to September in 2014, extreme climate change occurs, *ie* in the category of dry months (months with an average of less than 100 mm rainfall), resulting in drought or pneumonia area of 807 ha or higher than 54 ha in 2013. This dry season has an effect on the retreat of planting season and harvest season about one month. If you look at the total loss of the data, this amount is very large.

Therefore, the information regarding the pattern of rainfall is very useful for farmers in anticipating the possibility of extreme events which cause the failure of agricultural production. Many researchers studied the method for modeling the rainfall pattern. Piani et al. [4] planned and applied the bias correction on the output of daily climate model in Europe, so a distribution which approximates the observation model was obtained. This method is known as Statistical Bias Correction. The obtained result shows that the daily climate model is improved consistently with the method of bias correction. Bambang, et al. [5] identified the systematic error in TRMM data and built the transfer function to correct the bias of TRMM rainfall. From this research, it was found that the rainfall pattern from TRMM data which have been corrected and observed is similar, but the amount of rainfall is not similar.

In this research, Statistical Bias Correction method to see the relation between ECMWF (European Centre for Medium-Range Weather Forecasts) rainfall forecast and observation model from Indonesian Agency for Meteorological, Climatological, and Geophysics in 1996-2015 will be used. This research aims to identify the systematic error in ECMWF data and to build the transfer function in order to correct the up-to-date ECMWF rainfall bias as the prediction model. The systematic error can be identifiable through comparison of form and distribution of ECMWF hindcast data with observation data using the approaches of interpolation and the closest point. The resulted ensemble model in this research is the predicted rainfall approximation for the period of January - December 2016. Then will see how good the results based on observations of monthly rainfall in 2016.

2. Research Methodology

2.1. Data

The data used in this research are the rainfall data of ECMWF (*The European Centre for Medium-Range Weather Forecasts*) and observation in weather station from Central Indonesian Agency for Meteorological, Climatological and Geophysics in Jakarta from 1996 to 2016. The data taken is for observation in Bali Island.

2.2. Method of Data Analysis

2.2.1.1. Data Extraction. Data extraction was performed on the ECMWF rainfall data grid with MATLAB software to get the data output which is the same with the observation data. The data were extracted based on the location of latitude and longitude from every weather station. The process of data extraction used 2 methods, namely:

- *Method of the Closest Point.* It was performed by making an algorithm to determine the distance between 2 points and the closest distance. The determination of distance among the points $A(x_1, y_1)$ and $B(x_2, y_2)$ used the formula [6],

$$|AB| = ((x_2 - x_1)^2 - (y_2 - y_1)^2)^{1/2} \quad (1)$$

- *Cubic Interpolation.* Interpolation is a way of finding value among some known data points. This method is a numerical approach that needs to be done if we need the value of a function $y = y(x)$ which is not known its formulation appropriately. In this research, cubic

interpolation method will be used as one method for extracting ECMWF rainfall data. Cubic interpolation is done by determining the points between the four points by using the cubic polynomial approach as follows [7],

$$f_3(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)b_1(x - x_1) + b_3(x - x_0)(x - x_1)(x - x_2) \quad (2)$$

with,

$$b_0 = f(x_0) \quad (3)$$

$$b_1 = \frac{f(x_1) - f(x_0)}{(x_1 - x_0)} \rightarrow f[x_1, x_0] \quad (4)$$

$$b_2 = \frac{f[x_2, x_1] - f[x_1, x_0]}{(x_2 - x_0)} = \frac{\frac{f(x_2) - f(x_1)}{(x_2 - x_1)} - \frac{f(x_1) - f(x_0)}{(x_1 - x_0)}}{(x_2 - x_0)} \rightarrow f[x_2, x_1, x_0] \quad (5)$$

$$b_3 = \frac{f[x_3, x_2, x_1] - f[x_2, x_1, x_0]}{(x_3 - x_0)} \rightarrow f[x_3, x_2, x_1, x_0] \quad (6)$$

2.2.1.2. Identification of Data Distribution. In this stage, the parameter estimation process and the distribution matching which was the closest to the data were performed. This process was required to determine the cdf for the bias correction process. This data distribution was checked with the support of MATLAB software to match data with one of the fittest distributions, namely beta distribution, birnbaum-saunders, exponential, extreme value, gamma, generalized extreme value, generalized pareto, inverse Gaussian, logistic, log-logistic, lognormal, nakagami, normal, rayleigh, rician, t location-scale and weibull. To sort the fittest distribution, it was characterized based on the parameter as follows: NLogL (Negative of the log likelihood), BIC (Bayesian information criterion), AIC (Akaike information criterion), and AICc (AIC with a correction for finite sample sizes).

2.2.1.3. Bias Correction. Bias correction was performed to see the relation between data of the observed rain and the predicted rainfall (ECMWF data) to make certain transfer function [$y = f(x)$]. The form of relation is $cdf_{observed} = cdf_{predicted}$. The relation between the observed and predicted data can be in form of regression equation, which is linear, exponential, or polynomial. The cdf can be formed with the formula as follows:

$$cdf(x) = \int_0^x \frac{e^{\left(-\frac{x}{\theta}\right)} x'^{(k-1)}}{\Gamma(k)\theta^k} dx' + cdf(0) \quad (7)$$

where x = average daily rainfall,

k = parameter of form,

θ = scale, and

$\Gamma(.)$ = Gamma function, and this function can be evaluated with the factorial function

$$(\Gamma(k) = (k-1)!).$$

The cumulative distribution function (cdf) for data with zero value means that there is no rain on that day (*dry spell*) [4]. The result of this bias correction procedure was then used to correct model of the predicted rainfall in 2016.

2.2.1.4. Evaluation of the Best Prediction and Determination Models. The prediction evaluation was performed by considering and comparing the mean values and standard deviations from distribution of PDF rainfall model. The best model was determined by the following metric:

$$x = \frac{\bar{x}_{cor} - x_{obs}}{|\bar{x}_{fut} - x_{obs}|} \quad \text{and} \quad y = \frac{\sigma_{cor}}{\sigma_{fut}} \quad (8)$$

with,

$\bar{x}_{cor}$ = mean from pdf of the corrected rainfall,

$\bar{x}_{fut}$ = mean from pdf of rainfall before correction,

x_{obs} = the observed monthly rainfall,

σ_{cor} = standard deviation from pdf of the corrected rainfall, and

σ_{cor} = standard deviation from pdf of rainfall before correction.

The statistical bias correction perform well when the values metrics above are within a unit interval: $-1 < x < 1$ and $y < 1$ [8–9]

3. Results

3.1. General Description of Data

General description of the pattern of ECMWF rainfall data and the observed rainfall data from Indonesian Agency for Meteorological, Climatological and Geophysics is shown in Figure 1.

Figure 1. Overview of Rainfall Data Year 1996-2016.

(a) Plot of Rainfall Data Prediction of EMWF for Indonesia Areas.

(b) Rainfall Observation Data Plot of BMKG Observation for Sanglah Region in January.

Based on Figure 1(a), it can be observed that rainfall data from ECMWF is grid-shaped. Data Grid is the data presentation with interpolation method that is the value estimation process in a region which is not sampled or measured, so map or distribution of region is formed. The rainfall data have 4 parameters, namely: average daily rainfall, latitude, longitude, and time.

Figure 1(b) is the monthly rainfall data from the observation of Indonesian Agency for Meteorological, Climatological, and Geophysics for Sanglah station in January. These two data were taken in 1996-2016, in which 1996-2015 data will be used to see the relation between these two types of data and 2016 data are the basis for verification of the predicted rainfall.

3.2. Data Extraction

After the data extraction on the ECMWF rainfall data grid with MATLAB software, it resulted in 11 data of ensemble members for every observation month in 1996-2015. Meanwhile, from the process of data extraction in 2016, it resulted in 51 ensemble members data.

3.3. Identification of Data Distribution

The distribution determination which is the most suitable with the data is consideration of values of $NLogL$ (Negative of the log likelihood), BIC (Bayesian information criterion), AIC (Akaike information criterion), and $AICc$ (AIC with a correction for finite sample sizes) which is the smallest in Table 1.

Table 1. Checking of ECMWF Rainfall Data Distribution for Sanglah Region in January 1996-2015.

VAR (p)	NLogL	BIC	AIC	AICc
Weibull	1344.238	2699.264	2692.477	2692.532
Details	1344.858	2700.503	2693.716	2693.771
Nakagami	1345.544	2701.876	2695.088	2695.144
Normal	1346.4	2703.588	2696.801	2696.856
Generalized Extreme Value	1345.807	2707.795	2697.614	2697.725
t-Location Scale	1346.399	2708.98	2698.799	2698.91
Logistic	1349.854	2710.496	2703.709	2703.764
Gamma	1350.809	2712.406	2705.618	2705.674
Loglogistic	1358.745	2728.278	2721.491	2721.546
Lognormal	1362.047	2734.881	2728.093	2728.149
Birnbaumsaunders	1363.756	2738.299	2731.512	2731.567
Rayleigh	1367.05	2739.494	2736.1	2736.119
Inverse Gaussian	1365.038	2740.863	2734.076	2734.132
Extreme Value	1369.441	2749.669	2742.882	2742.937
Exponential	1473.633	2952.659	2949.265	2949.284

Based on the result in Table 1, ECMWF rainfall data for Sanglah Region in January 1996-2015 have Weibull distribution.

3.4. Bias Correction

In this research, the process of rainfall bias correction, as a step in making a new predicted rainfall model, was performed. ECMW rainfall model in 1996-2015 was corrected based on the observed rainfall in the same period.

This process was performed for every month in different location, so every location will result in the rainfall model from January to December 2016. The process of statistical bias correction was simulated like in Figure 2 below.

Figure 2. Process of statistical bias correction to compile rainfall model in January 2016 in Sanglah City.

- (a) Figure of pdf of the predicted rain (thick line) and the observed rain (thin line),
- (b) Figure of cdf value obtained from integral of pdf in Figure a,
- (c) Figure of cdf of rain model before correction (thick line) and the corrected rain model (thin line) was obtained from relation of $cdf_{observed}(y) = cdf_{predicted}(x)$ (thick line) in Figure b, and
- (d) Figure of pdf of rain model before correction (thick line) and rain prediction model after correction (thin line).

The result from bias correction process by considering the relation of $cdf_{observed}(y) = cdf_{predicted}(x)$ is then presented in PDF plot, so it is easier to be interpreted.

Figure 3. PDF of the Corrected Rainfall Model vs. Rainfall Model Before Correction in 2016 in Sanglah with Interpolation Approach.

Based on Figure 3, the bias correction method can identify well the bias of the historical rainfall forecast model. The value of the transfer function formed from the bias correction process can be used to improve the distribution of the 2016 rainfall prediction in Sanglah area. This can be seen from the rainfall prediction model in 2016 in Sanglah that has been improved based on information of rainfall bias from historical rainfall model.

3.5. The Best Prediction Model

After the precedures were applied, we can analyse the following result from the following table as follows:

Table 2. Evaluation of the Corrected Rainfall Model in 2016 in Sanglah with Interpolation Approach.

Month	$\bar{m}_{mod} - obs$	$\bar{m}_{cor} - obs$	σ_{mod}	σ_{cor}	$x = \frac{\bar{x}_{cor} - obs}{ \bar{x}_{mod} - obs }$	$y = \frac{\sigma_{cor}}{\sigma_{mod}}$
January	113.9	67.56	189.11	213.62	0.86	1.13
February	-216.45	-323.49	332.24	279.22	-1.49	0.84
March	145.922	63.56	226.98	176.32	0.44	0.78
April	120.16	28.49	163.53	96.31	0.24	0.59
May	149.46	85.23	193.09	131.79	0.57	0.68
June	-57.304	-105.24	73.91	22.43	-1.84	0.30
July	33.92	-1.57	116.55	62.73	-0.05	0.54
August	33.375	-15.08	76.45	24.19	-0.45	0.32
September	-150.91	-196.36	89.65	58.71	-1.30	0.65
October	37.52	-32.12	173.05	137.25	-0.86	0.79
November	-111.68	6.87	219.56	329.54	0.06	1.50
December	-56.60	-53.20	349.17	390.07	-0.94	1.12

Based on Table 2, there is information that the predicted rainfall model in Sanglah with interpolation approach is good for prediction in March, April, May, July, and October. It can be seen from x and y values which have met the criteria. In the months, x value is on range of $-1 < x < 1$ which means that mean from the corrected rainfall distribution is not far from the real value of rainfall realization. This value also implies that the accuracy of the new prediction model obtained is better than the prediction model before the rainfall bias is corrected. Then, y value which is less than 1 indicates that the actual rainfall, is estimated by an improved distribution with reduced standard deviation. This value also shows that the new prediction model obtained has better precision than the prediction model before the rainfall bias is corrected. For the cases in other months, the predicted rainfall model after correction can be said quite good since it meets one of the desired criteria.

Based on Table 3, the predicted rainfall after correction which is modeled with the closest point approach also gets a good result for some cases, in February and May. However, it is not for performing well in some months, such as September and December, which do not meet the determined criteria. With the x value which does not meet this criterion, it indicates that the mean value of the corrected rainfall distribution in these months is far from the real value of rainfall. Besides that, having a larger y value, indicates larger uncertainties where obtained after the procedure was applied. This value also indicates that the new prediction model obtained has a precision that is not better than the prediction model before it is corrected with its rainfall bias. So using this method is still less good when compared with using interpolation method.

Table 3. Evaluation of the Corrected Rainfall Model in 2016 in Sanglah with the Closest Point Approach.

Month	$\bar{m}_{mod} - obs$	$\bar{m}_{cor} - obs$	σ_{mod}	σ_{cor}	$x = \frac{\bar{x}_{cor} - obs}{ \bar{x}_{mod} - obs }$	$y = \frac{\sigma_{cor}}{\sigma_{mod}}$
January	50.63	51.12	173.21	208.33	1.01	1.20
February	167.89	107.83	294.19	279.69	0.64	0.95
March	78.63	19.86	203.36	176.91	0.25	0.87
April	24.11	-52.23	146.86	97.33	-2.17	0.66
May	68.35	-10.78	191.36	121.02	-0.16	0.63
June	20.87	-48.93	145.57	79.54	-2.35	0.55
July	-1.97	-45.81	120.38	62.34	-23.22	0.52
August	-42.62	-80.36	77.32	29.02	-1.89	0.38
September	-29.73	-77.12	89.54	55.42	-2.59	0.62
October	35.76	-48.29	159.67	133.76	-1.35	0.84
November	88.48	57.29	211.80	221.15	0.65	1.04
December	209.59	369.86	333.04	452.19	1.76	1.36

For modeling results in other areas of the island of Bali, generally we obtain improved results for rainfall data estimation using interpolation when compared to using the nearest point approach. The results of the evaluation of the bias correction process for all observation areas on the island of Bali can be seen in the graphs in Figure 4 and 5.

The average results obtained for all regions have met the criteria that have been determined. It is generally found that during the rainy season in February, March, November and December, the use of bias correction gives better results than other months for all areas of observation on the island of Bali. At least the corrected model for the season can meet one of the criteria in determining the best model that is seen from the mean value for the accuracy and standard deviation values for the precision level of the model [8-9]. Meanwhile, the results obtained by approaching the data using the nearest point can not explain well if viewed the results based on the season that occurred in the territory of Indonesia.

Figure 4. Graphic of Ratio Mean & Standard Deviation of Rainfall in Bali with Interpolation Approach in January – December 2016.

Figure 5. Graphic of Ratio Mean & Standard Deviation of Rainfall in Bali with the Closest Point Approach in January – December 2016.

4. Conclusion

The bias correction method can identify well the bias of the historical rainfall forecast model. The value of the transfer function formed from the bias correction process can be used to improve the distribution of the 2016 rainfall prediction on the island of Bali, to obtain better prediction. It is generally found that during the rainy season in February, March, November, and December, the use of bias correction provides better results than the dry season in modelling rainfall forecasts with data approach using interpolation. Whereas, the data approach using the nearest point in general has not provided good results for cases in the rainy or dry season. This can be caused of little information rainfall data used by researchers. For further research can use more rainfall data so that it can produce a better distribution.

References

- [1] Swarinoto YS, Sugiyono 2011 Utilization of Air Temperature and Air Humidity in Regression Equations for Simulation Monthly Rain Total Prediction in Bandar Lampung. *Journal of Meteorology and Geophysics* **12** 271–281
- [2] Aldrian E, M Karmini, Budiman 2011 Adaptation and Miti-gasi Climate Change in Indonesia *Climate Change and Air Quality Center of Indonesian Agency for Meteorological, Climatological and Geophysics* 67
- [3] Aldrian, E 2014 Understanding Climate Dynamics in The Indonesian Archipelago Countries as The National Resilience Modality *Indonesian Agency for Meteorological, Climatological and Geophysics* 20–21
- [4] Piani, C., J.O. Haerter, and E. Coppola 2010 Statistical Bias Correction for Daily Precipitation in Regional Climate Models Over Europe *J TheorApplClimatol* **99** 187–192
- [5] Bambang DD, Rizaldi B, Bambang P, and Yuli S 2013 Evaluation of Rainfall TRMM Using Statistics Bias Correction Approach *Land and Climate Journal* **38** 20–34
- [6] Purcell EJ, Varberg D, Rigdon SE 2003 *Calculus 8th Edition* (USA : Prentice Hall, Inc)
- [7] Cheney W, Kincaid D 2008 *Numerical Mathematics and Computing 7th Edition* (USA : Thomson Brooks/Cole)

- [8] JCGIM 2008 International Vocabulary of Metrology *Joint Committee for Guides in Metrology* **3** 21–22
- [9] Taylor JR 1999 *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements* (Colorado : University Science Books)